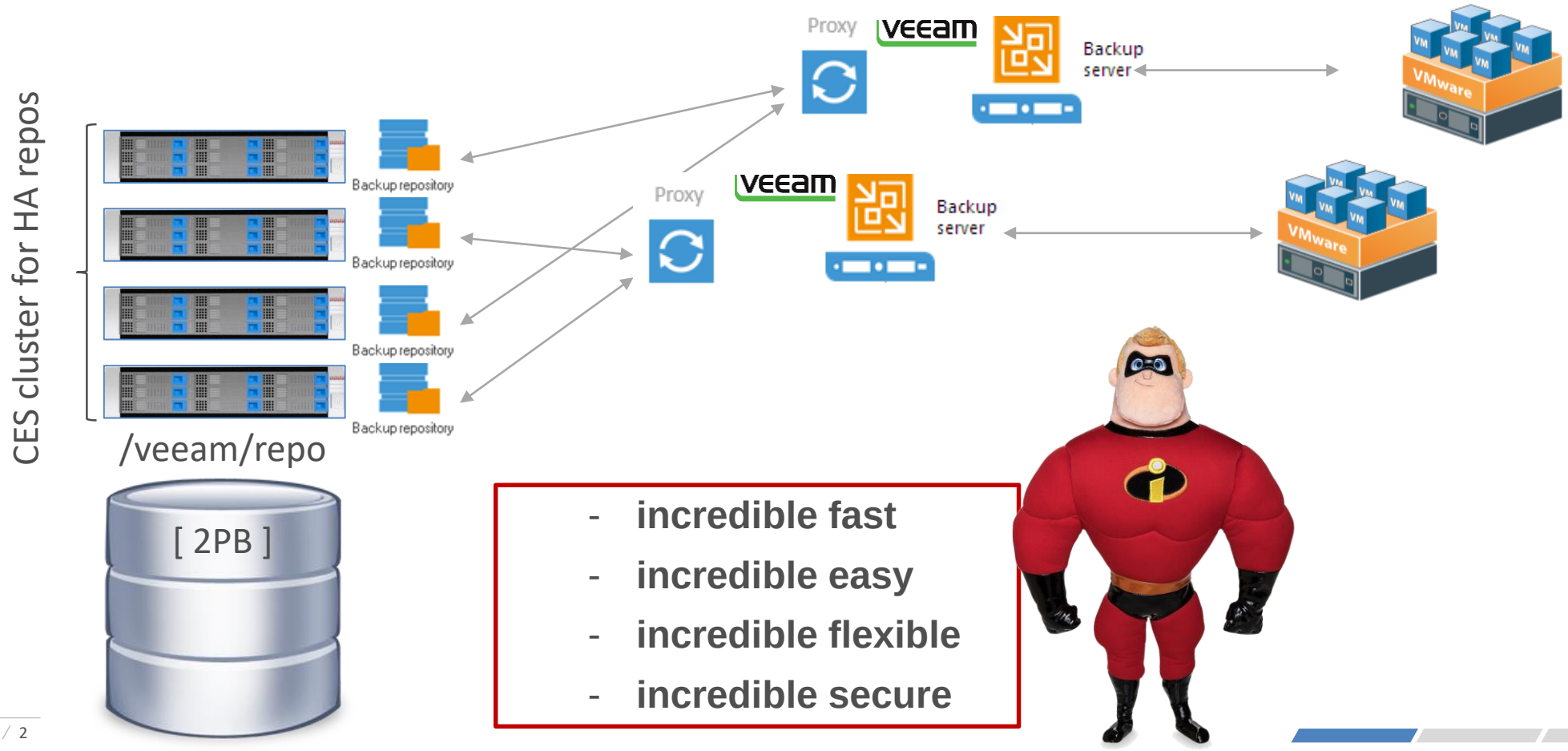
Small expenditure, large success - Spectrum Scale tips for the working day

jochen.zeller@sva.de

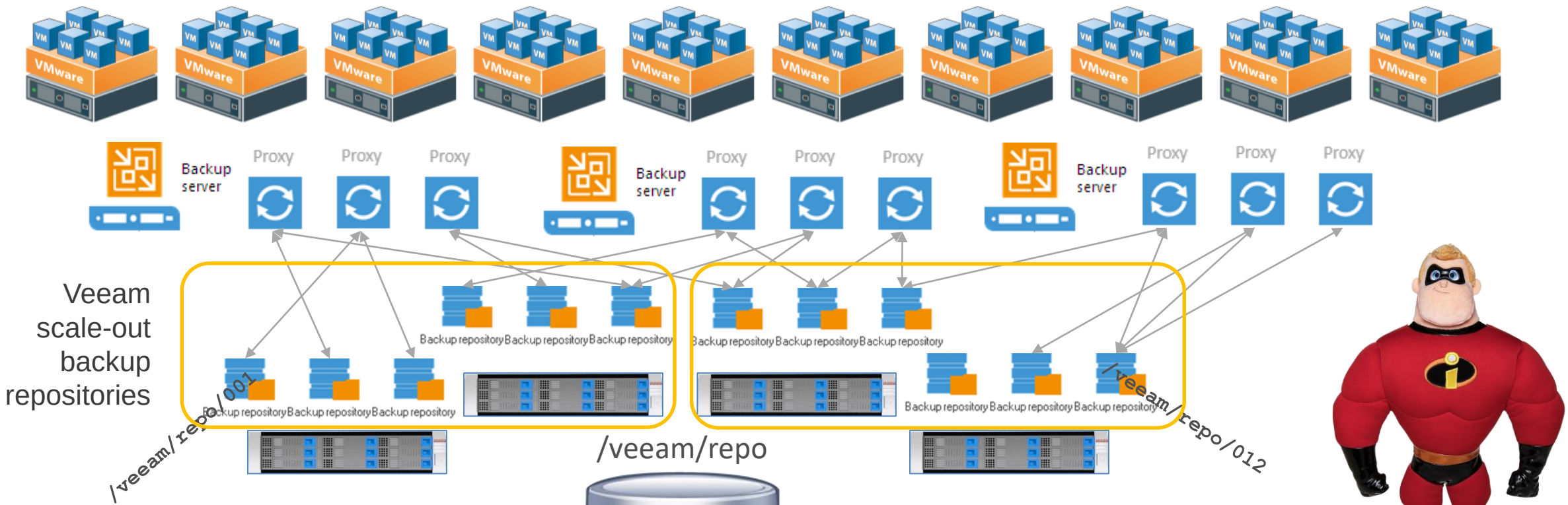
/ BUT FIRST, SOMETHING COMPLETELY DIFFERENT

Spectrum Scale as high available and high scalable repository for Veeam backups



/ BUT FIRST, SOMETHING COMPLETELY DIFFERENT

Spectrum Scale as high available and high scalable repository for Veeam backups



If you like further information about this solution, feel free to contact me!

SPECTRUM SCALE TIPS FOR THE WORKING DAY

SVA

/ MMHEALTH UND GUI EVENTS - HOW TO FIGHT

Already resolved errors that continue to be displayed in mmhealth and the GUI:

```
COMPONENT     NODE                                 STATUS            REASONS
-------------------------------------------------------------------------------
...
NODE          ENTERPRISE.UNIVERSE.COM              DEGRADED          PMSENSORS_DOWN
NODE          VOYAGER.UNIVERSE.COM                 DEGRADED          NETWORK_LINK_DOWN
NODE          DEFIANT.UNIVERSE.COM                 TIPS              GPFS_MAXFILESTOCACHE_SMALL
NODE          DEEPSPACE9.UNIVERSE.COM              DEGRADED          IB_RDMA_NIC_UNRECOGNIZED, NETWORK_LINK_DOWN
...
```

How to remove them (and this annoying TIPS):

```
mmdsh -N <NODE or all> mmsysmonc clearDB
mmdsh -N <NODE or all> mmsysmoncontrol restart
mmhealth event hide <EventName>
```

/ QOS IS COOL

QoS – Quality of Service gives you the opportunity to restrict the number of IOPS for a maintenance task.

- Enabling QoS:

```
# mmchqos <filesystem> --enable pool=system,maintenance=1000IOPS,other=unlimited
```

- This enables QoS for pool “system” and limits tasks with qos-class “maintenance” to 1000IOPS
- Use QoS in a maintenance command:

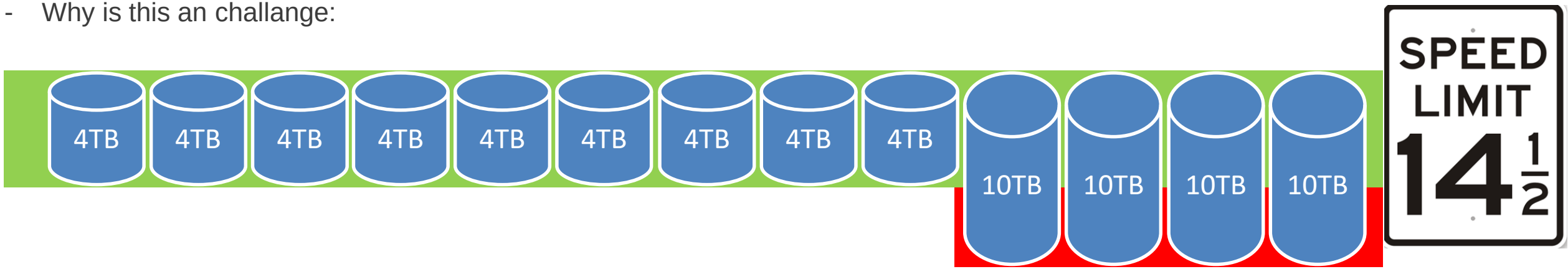
```
# mmdeldisk <filesystem> disk01 --qos maintenance -N nsd01,nds02
```

- When using multiple nodes in the command, than the IOPS are distributed among the nodes (in this case 500 + 500)

/ CLUSTER EXPANSION WITH LARGER HARD DRIVES

Typical situation: current cluster is running with e.g. 4TB NL-SAS drives, next expansion is with e.g. 10TB NL-SAS drives

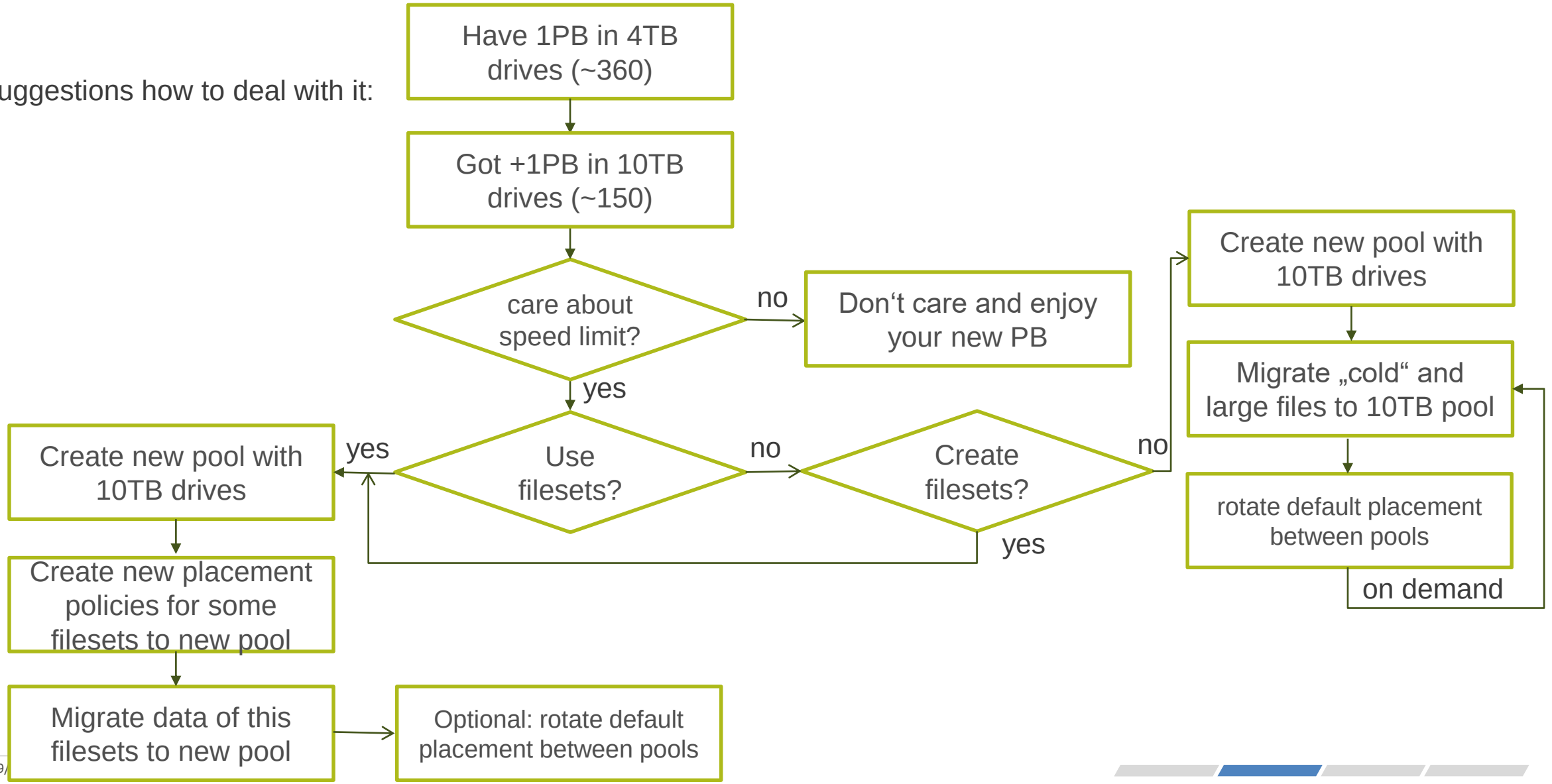
- Why is this an challange:



- When the areas with the 4TB disks are full, new data is only written to the 10TB drives - this significantly reduces the write and read performance for new data.

/ CLUSTER EXPANSION WITH LARGER HARD DRIVES

Some suggestions how to deal with it:



/ CLUSTER EXPANSION WITH LARGER HARD DRIVES

Some thoughts about filesets:

- A filesets could simplify the management of a project directory
  - With it's own inode space, the fileset is independent from the root fileset
  - Allows snapshots on a fileset base, independent from other filesets / filesystem snapshots
  - Fileset quotas and filesetdf could be used (shows with `df -h .` the hard limit capacity of the quota)
- Independent filesets use their own inode space, you must monitor and manage this number of inodes
  - mmhealth thresholds (+GUI +REST) monitors the usage:

```
# mmhealth thresholds list
### Threshold Rules ###
rule_name             metric                error  warn    direction filterBy  groupBy                                  sensitivity
-------------------------------------------------------------------------------------------------------------------------------
InodeCapUtil_Rule     Fileset_inode         90.0   80.0    high      gpfs_cluster_name,gpfs_fs_name,gpfs_fset_name      300
DataCapUtil_Rule      DataPool_capUtil      97.0   90.0    high      gpfs_cluster_name,gpfs_fs_name,gpfs_diskpool_name  300
...
```

/ CLUSTER EXPANSION WITH LARGER HARD DRIVES

Typical policy examples:

```
/* fileset placement project1 to new 10TB NL-SAS pool */
RULE 'project1' SET POOL 'NLSAS10TB' FOR FILESET ('project1')

/* fileset placement project2 to new 10TB NL-SAS pool */
RULE 'project2' SET POOL 'NLSAS10TB' FOR FILESET ('project2')

/ * default placement to pool NLSAS4TB, must be the last line */
RULE 'default' SET POOL NLSAS4TB'
```

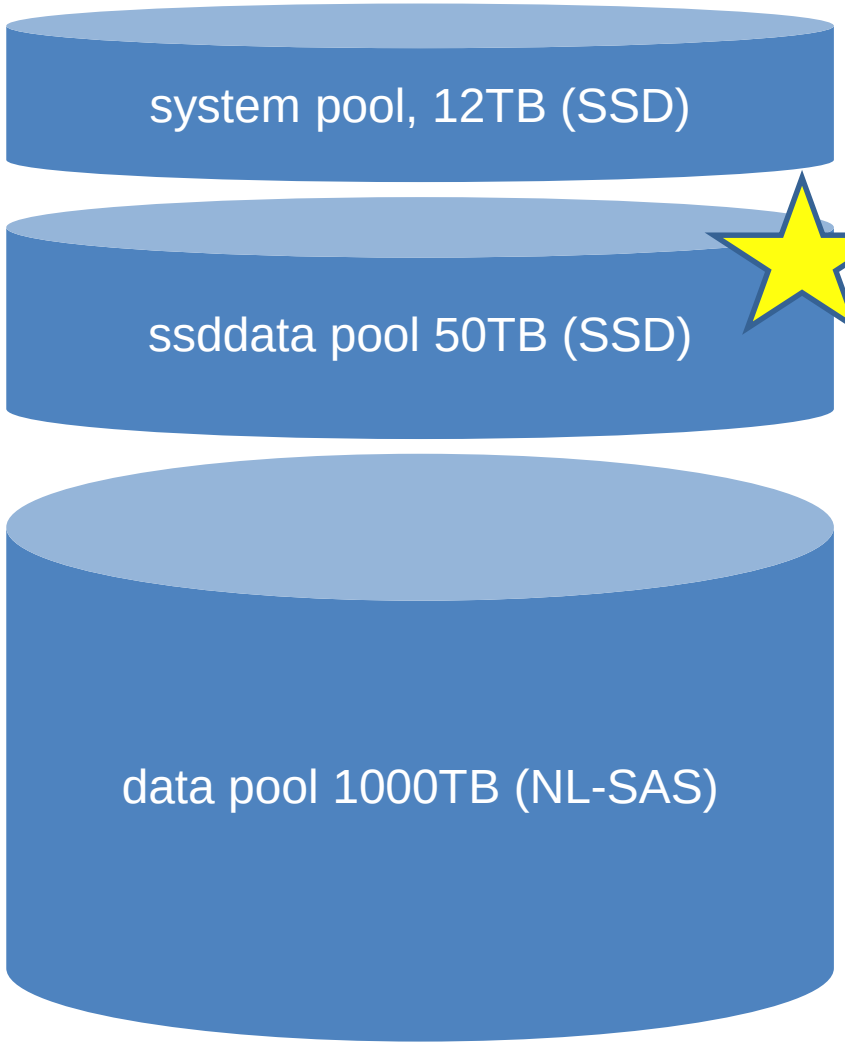
/ FOUND AN ALL-FLASH STORAGE – NO IDEA WHAT TO DO WITH

Example from a HPC project:

metadataOnly

default pool for all files

Applications / projects which mostly generates files >10GB should (must) point to data pool (by fileset rule)



Initial fill up:

migrate files ≤16M from pool data to pool ssddata (with QoS!)

daily: migrate files >128MB from pool ssddata to pool data (with QoS!)

threshold: if ssddata > 60% migrate files (to 50%) to pool data, weight by filesize (without QoS!)

/ FOUND AN ALL-FLASH STORAGE – NO IDEA WHAT TO DO WITH

What we achieved with this solution:

- Small files (we hope ≤16M) remain on SSD, performance of applications working with small files increased dramatically
- Most applications gain of the SSD read and write performance, even if they work with larger files on SSD
- The applications / projects separated to NL-SAS gain from their isolation, the available performance is less shared
- Large sequential processing jobs, separated to NL-SAS now, will no longer annoy (almost) all other users
- Overall IOPs + throughput is increased
- Hopefully we will never reach 100% usage on “ssddata”

/ CLUSTER EXPANSION WITH LARGER HARD DRIVES

Policy stuff behind:

- Extension of the placement policy by the threshold:

```
RULE 'migrate_SSD2NLSAS_60' MIGRATE FROM POOL 'ssddata' THRESHOLD(60,50)
 WEIGHT(KB_ALLOCATED) TO POOL 'data'
 WHERE ((CURRENT_TIMESTAMP - MODIFICATION_TIME) > INTERVAL '5' MINUTES)
```

- Daily migration policy started by cron on the file system manager, with QoS:

```
RULE 'migrate_SSD2NLSAS_128M' MIGRATE FROM POOL 'ssddata' TO POOL 'data' WHERE KB_ALLOCATED >
131072 AND ((CURRENT_TIMESTAMP - MODIFICATION_TIME) > INTERVAL '5' MINUTES)
```

- Add the callback to trigger the threshold migration on “lowDiskSpace”, without QoS:

```
mmchconfig enableLowspaceEvents=yes
/usr/lpp/mmfs/bin/mmaddcallback MIGRATION --command /usr/lpp/mmfs/bin/mmstartpolicy --event
lowDiskSpace --parms "%eventName %fsName -g /cfs/.policywdir/global -s /cfs/.policywdir/local -N
nsdsrv --single-instance --qos other"
```

/ THE END!

/ CLUSTER EXPANSION WITH LARGER HARD DRIVES

Policy stuff behind:

- Extension of the placement policy by the threshold:

```
RULE 'migrate_SSD2NLSAS_60' MIGRATE FROM POOL 'ssddata' THRESHOLD(60,50)
 WEIGHT(KB_ALLOCATED) TO POOL 'data'
 WHERE ((CURRENT_TIMESTAMP - MODIFICATION_TIME) > INTERVAL '5' MINUTES)
```

- Daily migration policy started by cron on the file system manager:

```
RULE 'migrate_SSD2NLSAS_128M' MIGRATE FROM POOL 'ssddata' TO POO
131072 AND ((CURRENT_TIMESTAMP - MODIFICATION_TIME) > INTERVAL '5
```

- Add the callback to trigger the threshold migration on "lowDiskSpace":

```
mmchconfig enableLowspaceEvents=yes
/usr/lpp/mmfs/bin/mmaddcallback MIGR      usr/lpp/mmfs/bin/mmstartpolicy --event
lowDiskSpace --parms "%eventName %fsN      olicywdir/global -s /cfs/.policywdir/local -N
nsdsrv --single-instance --qos other"
```

SVA 3/19/2019 / 13

Many thanks for your attention!

/ KONTAKT

JOCHEN ZELLER

System Architekt

Tel.: +49 151 180 256 77

Mail: jochen.zeller@sva.de